Cross-scene Crowd Counting via Deep Convolutional Neural Networks

Cong Zhang[1,2] Hongsheng Li[2,3] Xiaogang Wang[2] Xiaokang Yang[1]
[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2]Department of Electronic Engineering, The Chinese University of Hong Kong
[3]School of Electronic Engineering, University of Electronic Science and Technology of China

Counting crowd pedestrians in videos draws a lot of attention because of its intense demands in video surveillance, and it is especially important for metropolis security. Crowd counting is a challenging task due to severe occlusions, scene perspective distortions and diverse crowd distributions. Since pedestrian detection and tracking has difficulty being used in crowd scenes, most state-of-the-art methods [1, 2, 3, 4, 5, 6, 7] are regression based and the goal is to learn a mapping between low-level features and crowd counts. However, these works are scene-specific, i.e., a crowd counting model learned for a particular scene can only be applied to the same scene. Given an unseen scene or a changed scene layout, the model has to be re-trained with new annotations. There are few works focusing on cross-scene crowd counting, though it is important to actual applications.

In this paper, we propose a framework for cross-scene crowd counting. No extra annotations are needed for a new target scene. Our goal is to learn a mapping from images to crowd counts, and then to use the mapping in unseen target scenes for cross-scene crowd counting. To achieve this goal, we need to overcome the following challenges. 1) Develop effective features to describe crowd. Previous works used general hand-crafted features. These features have low representation capability for crowd. New descriptors specially designed or learned for crowd scenes are needed. 2) Different scenes have different perspective distortions, crowd distributions and lighting conditions. Without additional training data, the model trained in one specific scene has difficulty being used for other scenes. 3) For most recent works, foreground segmentation is indispensable for crowd counting. But crowd segmentation is a challenging problem and can not be accurately obtained in most crowded scenes. The scene may have stationary crowd without movement. 4) Existing crowd counting datasets are not sufficient to support and evaluate cross-scene counting research. The largest one [5] only contains 50 static images from different crowd scenes collected from Flickr. The widely used UCSD dataset [1] and the Mall dataset [2] only consist of video clips collected from one or two scenes.

Considering these challenges, we propose a Convolutional Neural Network (CNN) based framework for cross-scene crowd counting. CNN models have strong discriminative capability and can represent several patterns with visual characteristics disparity. After a CNN is trained with a fixed dataset, a data-driven method is introduced to fine-tune (adapt) the learned CNN to an unseen target scene, where training samples similar to the target scene are retrieved from the training scenes for fine-tuning. Figure 1 illustrates the overall framework of our proposed method. Our cross-scene crowd density estimation and counting framework has following advantages:

1. Our CNN model is trained for crowd scenes by a switchable learning process with two learning objectives, crowd density map and crowd counts. The two different but related objectives can alternatively assist each other to obtain better local optima. Our CNN model learns crowd-specific features, which are more effective and robust than handcrafted features. An overview of our crowd CNN model with switchable objectives is shown in Fig 2.

2. The target scenes require no extra labels in our framework for cross-scene counting. The pre-trained CNN model is fine-tuned for each target scene to overcome the domain gap between different scenes. The fine-tuned model is specifically adapted to the new target scene.

3. The framework does not rely on foreground segmentation results because only appearance information is considered in our method. No matter whether the crowd is moving or not, the crowd texture would be captured by the CNN model and can obtain a reasonable counting result.

4. We also introduce a new dataset for evaluating cross-scene crowd counting methods. The dataset covers a large variety of scenes and crowd distributions. It contains 108 different crowd scenes with nearly 220,000 pedestrian annotations. To the best of our knowledge, this is the largest dataset for evaluating crowd counting algorithms. Extensive experiments on the proposed and another two existing datasets [1, 5] demonstrate the effectiveness and reliability of our approach.

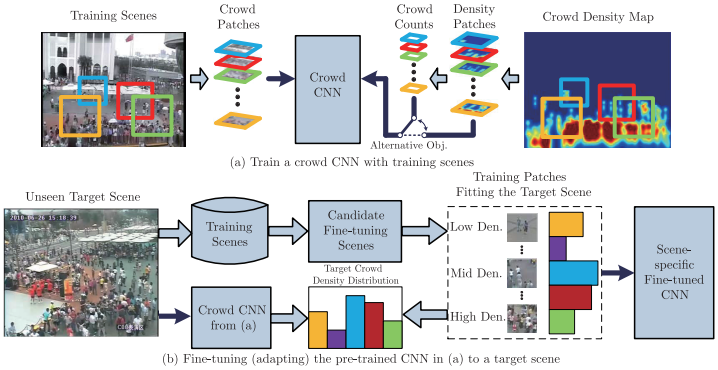

Figure 1: Illustration of our proposed cross-scene crowd counting method.

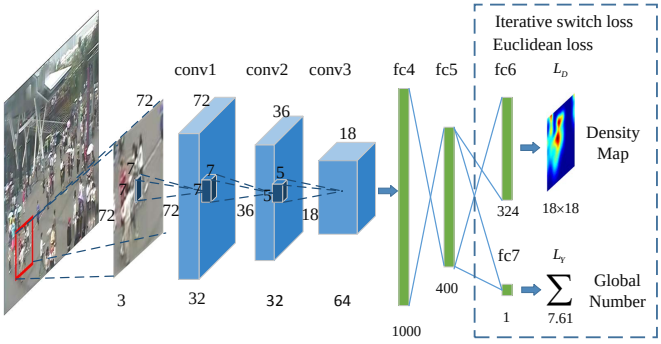

Figure 2: The structure of the crowd convolutional neural network. At the loss layer, a density map loss and a global count loss is minimized alternatively.

[1] A. B. Chan, Z. S. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008.

[2] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *BMVC*, 2012.

[3] Ke Chen, Shaogang Gong, Tao Xiang, Queen Mary, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *CVPR*, 2013.

[4] Luca Fiaschi, Rahul Nair, Ullrich Koethe, and Fred A Hamprecht. Learning to count with regression forest and structured labels. In *ICPR*, 2012.

[5] H. Idrees, I. Saleemi, and M. Shah. Multi-source multi-scale counting in extremely dense crowd images. In *CVPR*, 2013.

[6] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *NIPS*, 2010.

[7] C. C. Loy, S. Gong, and T. Xiang. From semi-supervised to transfer counting of crowds. In *ICCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.